Conselho Nacional de Arquivos

Câmara Técnica de Documentos Eletrônicos

Orientação Técnica nº 4

Outubro de 2016

Recomendações de uso do PDF/A para Documentos Arquivísticos

ESCOPO DA ORIENTAÇÃO TÉCNICA

Esta orientação técnica apresenta recomendações gerais sobre o uso do formato PDF/A na produção e no arquivamento de documentos arquivísticos digitais, ou seja, nas idades corrente, intermediária e permanente, visando o seu acesso e a sua preservação. Os objetivos desta nota técnica são:

a) **esclarecer as principais características do PDF/A e seus subtipos;**
b) **apresentar alguns cenários de uso;**
c) **ressaltar os benefícios, limitações e riscos do uso desse formato.**

Não está no escopo desta orientação a apresentação técnica exaustiva do PDF/A. Para isso, deve-se consultar as normas técnicas publicadas pela ISO 19005 (ISO 19005-1:2005, ISO 19005-2:2011, ISO 19005-3:2012). Aqui serão apresentados apenas os principais conceitos que contribuem para o entendimento das recomendações pela comunidade arquivística nos diversos cenários de uso.

INTRODUÇÃO

A preservação e o acesso aos documentos digitais dependem de uma série de cuidados que devem ser considerados sob pena de comprometer a autenticidade, o acesso e o uso desses documentos ao longo do tempo. Um desses cuidados é a definição de formatos de arquivo com características que possam permitir a preservação e o acesso de documentos digitais, com independência de sistemas operacionais e hardware.

Nesse sentido, desde o ano de 2005, o formato de arquivo denominado PDF/A (*Portable Document Format/Archive*) foi instituído como norma ISO (*International Organization for Standardization*) para preservação de alguns tipos de documentos digitais em longo prazo. A especificação do PDF/A como norma ISO torna o formato um padrão aberto, tornando-se de amplo uso e facilitando a criação de aplicativos pelos desenvolvedores.

O formato PDF/A atende à produção dos documentos textuais e imagéticos paginados, permitindo manter sua forma fixa e conteúdo estável.

As características do PDF/A tem incentivado o seu crescente uso pelas organizações, governos e pessoas. No Brasil, o PDF/A é um dos formatos de arquivo adotados pelo governo federal, previsto nos Padrões de Interoperabilidade de Governo Eletrônico (e-PING).

A Câmara Técnica de Documentos Eletrônicos – CTDE, no intuito de orientar os órgãos e entidades integrantes do Sistema Nacional de Arquivos – SINAR/CONARQ, vem por meio desse documento, fornecer subsídios para a melhor utilização do PDF/A.

I) BREVE HISTÓRICO DO PDF E DO PDF/A

Originalmente, o formato PDF foi criado em 1993, pela empresa Adobe Systems@,[1] como um formato de documento orientado à página,[2] compacto e com a capacidade de manter a visualização original independente de plataforma (hardware e sistema operacional), podendo ser criado a partir de outros formatos digitais, e assim tornou-se rapidamente um padrão *de facto*.[3] Uma das outras razões de seu amplo uso é a possibilidade de desenvolvedores criarem e distribuírem conversores e leitores para PDF, sem, no entanto, poderem alterar as especificações originais da Adobe.

Nos treze anos seguintes, a empresa Adobe lançou oito versões que incluíram novas funcionalidades para os seus usuários. Algumas dessas novidades, no entanto, trouxeram características que dificultavam a preservação do documento. Por exemplo, o uso de arquivos de fontes tipográficas externas (*font linking*), ao mesmo tempo em que torna os arquivos menores, cria uma dependência de localização e acesso a fontes externas para a visualização do documento na sua forma original. Caso o arquivo da fonte tipográfica não esteja presente no computador em que se visualizava o arquivo PDF, o resultado apresentado em tela poderia ser diferente daquele pretendido pelo autor do documento. Para contornar esse tipo de problema, foi criado, em 2002, um grupo de trabalho na ISO para definição de um formato digital voltado especificamente para a preservação de documentos digitais em longo prazo, e que seria disponibilizado como uma especificação normalizada por entidade independente da Adobe Systems@.

Como resultado desse trabalho, em 1º de outubro de 2005, foi publicada a norma ISO 19005-1:2005 denominada *Document Management – Electronic document file format for long term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*.

Após a publicação da norma ISO 19005, baseada no PDF 1.4, a Adobe seguiu desenvolvendo novas versões do PDF como formato proprietário da empresa. Em 1º de julho de 2008, foi publicada a norma ISO 32000-1, equivalente à especificação PDF 1.7 da Adobe Systems. Dessa maneira, o formato PDF passou a ser uma norma técnica internacional, um padrão de direito (*de jure*), e suas futuras versões passaram a ser definidas pela ISO.

II) SUBTIPOS DO PDF/A

[1] Adobe Systems Inc. Disponível em: <http://www.adobe.com/> Acesso em: 26 nov. 2015.
[2] Orientado a página, significa seu uso para documentos multipáginas.
[3] Padrão de *facto*. São aqueles que se consagraram voluntariamente, devido as suas características consideradas de boa utilização. Independem de organizações normalizadoras, que são as que emitem os padrões *de jure*.

O formato PDF/A, até o momento, apresenta-se em três subtipos: PDF/A-1, PDF/A-2 e PDF/A-3. Estes subtipos não se constituem em especificações evolutivas do PDF/A e não substituem necessariamente as anteriores, mas possuem características próprias para atender determinadas finalidades.

a) PDF/A-1

Basicamente, o formato PDF/A-1 é uma "versão simplificada" da versão 1.4 do formato PDF da empresa Adobe Systems, na medida em que se proíbe uma série de características que dificultam a tarefa de preservação digital, tais como: códigos executáveis *javascript*, hyperlinks externos, inserção de áudio e vídeo. Por outro lado, o PDF/A-1 obriga que outras características, as quais facilitam a preservação digital, estejam presentes, tais como metadados e fontes embutidas (*embedding font*).

b) PDF/A-2

Em 2011, foi publicada a ISO 19005-2:2011 (Parte 2), que estabeleceu o formato PDF/A-2. Essa parte não substitui a anterior, apenas define um novo formato que considera novas características decorrentes da evolução do formato PDF ou simplesmente proibidas pela Parte 1. Por exemplo, o PDF/A-2 pode conter, como anexos, arquivos no formato PDF/A. Além disso, passaram a ser permitidas as seguintes características: transparência, camadas, compressão JPEG 2000 e assinatura digital avançada (PaDes).[4] De forma diferente da Parte 1, que era baseada no formato proprietário PDF 1.4, a Parte 2 é baseada na especificação ISO 32.000-1:2008.

c) PDF/A-3

Em outubro de 2012, foi publicada a ISO 19005-3:2012, *Document management — Electronic document file format for long-term preservation — Part 3: Use of ISO 32000-1:2008 with support for embedded files (PDF/A-3)*. Essa parte não substitui as anteriores, nem representa uma evolução do formato, apenas define um novo formato, que possibilita inserção de anexos em qualquer formato, inclusive diferente do PDF/A. Esses anexos, denominados ”arquivos associados” (*Associated Files*),

[4] *PDF Advanced Electronic Signatures* - PaDES. ETSI/Technical Specifcation (TS) 102 778. Essa especificação é para permitir a incorporação da assinatura digital avançada no documento em PDF. O Brasil é aderente aos padrões emitidos pela European Telecommunications Standards Institute – ETSI. Disponível em: <http://www.etsi.org/technologies-clusters/technologies/security/electronic-signature/>. Acesso em: 30 set. 2016. O Instituto de Tecnologia da Informação/ITI, regulamentou no Brasil, a aderência no âmbito da ICP-Brasil, a especificação PaDES por meio da Resolução n. 109/2015. Disponível em: <http://www.iti.gov.br/images/legislacao/resolucoes/resolucao_109.pdf> Acesso em: 30 set. 2016.

podem ser utilizados para armazenar dados, visualizações alternativas ou o próprio arquivo fonte que deu origem ao PDF/A-3. Em outras palavras, essa versão permite embutir no PDF/A-3 os arquivos que lhe deram origem em seus formatos nativos como DOC e XLS. Dessa maneira permite-se o envio de documentos reutilizáveis ou editáveis, necessários à realização de atividades dentro de um fluxo de trabalho. Ressalte-se que no caso dos arquivos associados em formatos diferentes de PDF/A, não existe expectativa de preservação em longo prazo.

As três partes da especificação PDF/A definem formatos de arquivos para a preservação digital que coexistem. A norma técnica define que um programa que visualize arquivos no formato PDF/A-3 ou PDF/A-2 deve, obrigatoriamente, visualizar os arquivos nos formatos PDF/A-1 e PDF/A-2. Sendo assim, caso uma organização possua arquivos em conformidade com o formato PDF/A-1, não é necessário convertê-los para os formatos PDF/A-2 ou PDF/A-3. Como a conversão de formatos de arquivos é realizada por software, existiria a possibilidade de ocorrer perda de informação nesse processo.

Cabe esclarecer que o uso de arquivos associados no PDF/A-3 traz preocupações quanto à preservação em longo prazo e mesmo quanto ao uso primário na segurança da informação. Nesse sentido, o relatório técnico *The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions* da *National Digital Stewardship Alliance* (NDSA),[5] de 2014, apresenta cenários hipotéticos em que o uso do formato PDF/A-3 pode ser benéfico ou apresentar riscos. De forma resumida e adaptada, seguem alguns possíveis cenários para a utilização do PDF/A-3:

- Inclusão de dados de pesquisa em documentos acadêmicos: utiliza-se o perfil *Data* em formatos abertos como CSV[6] ou XML.[7] Os dados podem ser relacionados a um elemento do documento (gráfico, tabela etc.) ou ao documento como um todo.
- Inclusão de arquivos de aplicações CAD:[8] utiliza-se o perfil *Source* para permitir a edição do documento original.
- Captura de documentos disponíveis na WEB com licença *Creative Commons*[9] que estão no formato PDF: a conversão para PDF/A embute o documento PDF original como arquivo associado utilizando o perfil *Source*, bem como recursos que garantem a preservação, tais como, fontes tipográficas e esquema de cores.
- Utilização do PDF/A combinado com XML em um fluxo de autoria: o PDF/A-3 apresenta a informação formatada, para conveniência de

[5] *The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions - National Digital Stewardship Alliance* (NDSA). EUA. 2014. Disponível em: <http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014>. Acesso em: 26 nov. 2015.

[6] CSV. Comma Separated Values (CSV) é um formato de arquivo de texto que pode ser usado para trocar dados de uma planilha entre aplicativos.

[7] XML. Extensible Markup Language (XML). Linguagem de marcação.

[8] CAD. Computer-Aided Design. São sistemas computacionais projetados para atividades de projetos de engenharia, arquitetura, geografia, geologia e *design* de produtos.

[9] Creative Commons. Um tipo de declaração de licença de uso e compartilhamento de recursos digitais (obras intelectuais, softwares). Disponível em: <creativecommons.org.br>. Acesso em: 26 nov 2015.

leitura, e embute a informação estruturada (XML), para facilidade de criação de novas versões.

III) PRINCIPAIS CENÁRIOS DE USO PARA DOCUMENTOS ARQUIVÍSTICOS

Uso em idade corrente e intermediária:

Os documentos podem ser produzidos diretamente em PDF/A ou em outros formatos e convertidos para PDF/A no momento do arquivamento.

No caso de documentos produzidos e mantidos na forma de tabelas de uma base de dados, o sistema informatizado poderá produzir um PDF/A para apresentar o documento ao usuário ou para exportá-lo para um sistema informatizado de gestão arquivística de documentos.

Uso em idade permanente:

O PDF/A pode ser utilizado para facilitar a preservação e o acesso a documentos digitais (originais digitais ou representantes digitais).[10]

No momento do recolhimento ao arquivo permanente, pode-se converter os documentos originais digitais para o formato PDF/A, de maneira a padronizar o formato no arquivo permanente e a dar mais garantias de acesso em longo prazo.

Em relação aos representantes digitais, podem-se produzir cópias de acesso em PDF/A. Os arquivos originais de imagem devem ser preservados no formato em que foram gerados.

Em ambos os casos, o documento original (digital ou não digital) deve ser mantido.

[10] Representante digital (*digital surrogate*): Imagem digital de um documento originalmente não digital, resultante da digitalização do mesmo com o uso de equipamentos como scanner e câmera.

IV) RECOMENDAÇÕES GERAIS

A seguir apresentamos algumas recomendações para a utilização do PDF/A e seus diversos subtipos.

RECOMENDAÇÃO 1

Utilize preferencialmente o formato PDF/A-1 para arquivamento de documentos.

Justificativa: O PDF/A-1 é o subtipo mais simples para geração, validação e visualização dos documentos.

RECOMENDAÇÃO 2

No caso da necessidade de utilização de assinatura digital avançada (PADES), use preferencialmente o formato PDF/A-2 para arquivamento de documentos.

Justificativa: O PDF/A-1 só permite a assinatura digital simples.

RECOMENDAÇÃO 3

Não converter arquivos PDF/A-1 para PDF/A-2 ou PDF/A-3, como também, não converter arquivos PDF/A-2 para PDF/A-3.

Justificativa: Essa conversão se mostra na prática desnecessária. O processo de conversão pode gerar perda de informações.

RECOMENDAÇÃO 4

Utilize o formato PDF/A-2 se for necessário anexar arquivos em formato PDF/A-1.

Justificativa: A verificação de arquivos no formato PDF/A-2 é mais simples de ser realizada do que no formato PDF/A-3.

RECOMENDAÇÃO 5

Verifique as licenças de fontes tipográficas embutidas no arquivo PDF/A.

Justificativa: O uso não autorizado de fontes tipográficas infringe direitos autorais.

RECOMENDAÇÃO 6

Utilizar o PDF/A para derivadas de acesso no caso de documentos digitalizados (obtidos de documentos não digitais) e preservar as matrizes digitais (preferencialmente produzidas em alta resolução).[11]

Justificativa: O PDF ou PDF/A (e suas versões) não são formatos de preservação de imagens. Além do mais, a guarda das matrizes em alta resolução facilita a reutilização das imagens originais.

RECOMENDAÇÃO 7

Ao utilizar o PDF/A para acesso e visualização de documentos digitalizados (obtidos de documentos não digitais), criar um arquivo PDF específico para cada documento (com uma ou mais páginas), evitando agregar documentos que não tenham relação entre si.

Justificativa: A criação de falsas agregações dificulta a administração, a pesquisa e a recuperação dos documentos.

RECOMENDAÇÃO 8

O PDF/A pode ser utilizado para normalização de formatos[12] no arquivamento de alguns tipos de documentos produzidos em outros formatos.

Justificativa: O PDF/A é uma norma publicada e especificada pela ISO, oferecendo uma expectativa de longo prazo para preservação e acesso dos documentos.

RECOMENDAÇÃO 9

Ao utilizar o formato PDF/A-3, defina o perfil de uso, especificando os tipos de arquivos autorizados para cada tipo de arquivo associado.

[11] Resolução de imagem é o nível de detalhe que uma imagem apresenta. Em imagens digitais se usam as medidas em pixels.

[12] Normalização de formatos. Conversão de formatos de arquivo para um elenco gerenciável de formatos apropriados para preservação e acesso.

Justificativa: O uso indiscriminado de arquivos associados pode trazer anexos não desejados, como nos casos de arquivos binários executáveis infectados por vírus.

RECOMENDAÇÃO 10

Quando do uso do PDF/A-3, não aceitar arquivos anexados que utilizem o perfil de associação ”Não Especificado” (*Unspecified*).

Justificativa: A descrição do tipo de associação é um metadado descritivo importante para a instituição arquivística.

RECOMENDAÇÃO 11

Quando do uso do PDF/A-3, não aceitar arquivos anexados que utilizem o ”MIME type” ”*application/octet-stream*”.

Justificativa: O uso de ”MIME type” genérico pode trazer conteúdos indesejados.

RECOMENDAÇÃO 12

Quando do uso do formato PDF/A-3 em documentos correntes, sempre verificar a existência de vírus nos arquivos associados no momento da entrada no repositório.

Justificativa: A possibilidade da existência de vírus representa risco para o repositório.

RECOMENDAÇÃO 13

Utilizar um validador de conformidade do formato PDF/A.

Justificativa: A conformidade do arquivo PDF/A é garantida pelo validador.

V) REFERÊRENCIAS TÉCNICAS

- *PDF/A in a natural Nutshell 2.0*. PDF Association. Disponível em: http://www.pdfa.org/publication. Acesso em: 30 set. 2016.
- *The Benefits and Risks of the PDF/A-3 File Format for Archival*

Institutions. National Digital Stewardship Alliance (NDSA). EUA. 2014. Disponível em: <http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014>. Acesso em: 30 set. 2016.

- Estados Unidos da América. *Sustainability of Digital Formats Libray of Congress. Planning for Library of Congress Collections*. Disponível em:<http://www.digitalpreservation.gov/formats/index.shtml> Acesso em: 30 set. 2016.

ANEXO

Nível de conformidade

A norma ISO 19.005 define três níveis de conformidade, como apresentado a seguir:

- No nível básico (*basic*: PDF/A-1b, PDF/A-2b e PDF/A-3b), garante-se a reprodução confiável da aparência visual do documento;
- No nível intermediário (PDF/A-2u e PDF/A-3u), define-se um nível que acrescenta ao nível básico a utilização do conjunto de caracteres Unicode;
- No nível de acessibilidade (*accessible*: PDF/A-1a, PDF/A-2a e PDF/A-3a), a norma define características que facilitam a acessibilidade e permitiriam, por exemplo, que um software reproduzisse, por meio de síntese de voz, o documento para deficientes visuais. Para que esse formato represente a estrutura lógica do documento e a ordem natural de leitura, utilizam-se marcas (*Tagged*), especificação de idioma e o conjunto de caracteres Unicode.

Fontes tipográficas

A inclusão das fontes tipográficas no próprio arquivo PDF/A possibilita a reprodução confiável da aparência visual do documento. Essa inclusão pode ser feita de forma completa ou parcial. No caso de inclusão parcial (*subset*), são incluídas apenas as definições dos caracteres utilizados pelo documento, diminuindo, dessa forma, o tamanho do arquivo final. Essa característica é importante, pois algumas fontes tipográficas podem possuir arquivos muito grandes.

A maioria das fontes tipográficas são recursos protegidos por direitos de uso. As licenças das fontes, normalmente, definem se elas podem ser embutidas (*embedding*) ou não embutidas (*non embedding*) no documento. No caso de permitir que seja embutida a fonte, ainda existe a especificação de se essa inclusão é apenas para visualização e impressão (*View and Print*), para edição (*Editable*) ou para instalação (*Installable*).

Programa leitor de PDF/A

No intuito de garantir em longo prazo o acesso, a norma 19005-1:2005, define requisitos a serem atendidos por programas visualizadores de arquivos PDF/A, desenvolvidos no tempo presente ou no futuro, dentre os quais destacamos:

- considerar apenas as fontes embutidas no arquivo;
- mostrar imagens utilizando a configuração (*profile*) de cores definida no arquivo;
- desabilitar modificações;
- apresentar anexos que estejam no formato PDF/A; e
- permitir a exportação de anexos que estejam em outros formatos.

Arquivos anexados (*Associated Files*)

O PDF/A-2, permite a inserção de arquivos PDF/A, e o PDF/A-3 expandiu a funcionalidade do padrão, na medida em que passou a permitir que o formato fosse utilizado também como um *container,* ou seja, permite que arquivos em outros formatos diferentes do padrão PDF, sejam incluídos como anexos.

No entanto, apenas os arquivos associados no formato PDF/A possuem garantia de preservação permanente.

Os arquivos associados podem ser relacionados com o documento como um todo, ou com uma parte dele, como uma tabela, uma imagem ou uma página. Para cada arquivo associado, deve ser definida uma descrição informando o objeto relacionado, o *mimetype*[13] e a natureza do relacionamento, que considera os seguintes tipos de componentes: arquivo fonte (*source*), arquivo de dados (*data*), arquivo alternativo (*alternative*), arquivo suplemento (*suplement*) e arquivo de natureza não especificada (*unspecified*).

[13] É o identificador padrão da Internet que indica o conteúdo de dados que o arquivo contém. Ex.: text – html; image – tiff.